THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 287, NO. 1, pp. 35–42, January 2, 2012

Inference of Functional Properties from Large-scale Analysis of Enzyme Superfamilies*

Published, JBC Papers in Press, November 8, 2011, DOI 10.1074/jbc.R111.283408

Shoshana D. Brown[‡] **and Patricia C. Babbitt**[‡§¶1]

From the Departments of [‡]*Bioengineering and Therapeutic Sciences and* [§]*Pharmaceutical Chemistry, School of Pharmacy, and* [¶]*California Institute for Quantitative Biosciences, University of California, San Francisco, California 94158-2330*

As increasingly large amounts of data from genome and other sequencing projects become available, new approaches are needed to determine the functions of the proteins these genes encode. We show how large-scale computational analysis can help to address this challenge by linking functional information to sequence and structural similarities using protein similarity networks. Network analyses using three functionally diverse enzyme superfamilies illustrate the use of these approaches for facile updating and comparison of available structures for a large superfamily, for creation of functional hypotheses for metagenomic sequences, and to summarize the limits of our functional knowledge about even well studied superfamilies.

In the post-genomic era, access to large amounts of gene sequence and protein structure data has become the norm; by mid-2011, the number of protein sequences in the UniProt/TrEMBL Database (1) topped 16 million, whereas the Protein Data Bank (2) contained over 73,000 structures. Additional millions of sequences are becoming available from newer types of genome projects, including metagenomics projects, with one report for the human gut microbiome accounting for an additional 3.3 million microbial genes (3). Because experimental determination of protein function lags far behind the rate of sequence and structure determination, improved computational methods for function prediction are urgently needed to help bridge the gap between sequenced genes and functionally characterized protein products. In response, new methods are rapidly being developed to address these challenges, and community efforts are now under way to increase the pace of experimental and computational prediction of protein function (4, 5). Another large-scale effort (http://www.nigms.nih.gov/News/Results/gluegrant_051510.htm) aims to develop a combined experimental/computational strategy for the prediction of the reaction and substrate specificity of enzymes, the protein class that is the subject of this minireview. Additionally, community challenges such as the Critical Assessment of Function Annotations (CAFA) (Automated Function Prediction 2011) have been mounted to assess and improve the current state of automated prediction of protein function. Viewing the glass as half-full, progress in sequencing and annotation over the last decade led one group to estimate that some functional features can be assigned to as much as 85% of proteins in completely sequenced genomes (6). From a more skeptical perspective, more recent assessments of annotation accuracy suggest that computational approaches are especially prone to misannotation (7, 8), indicating that significant challenges for functional inference remain.

This minireview focuses on how new insights about protein structure-function relationships and functional inference can be obtained from large-scale analyses of proteins, specifically for "functionally diverse" enzyme superfamilies. We define these types of superfamilies as sets of homologous proteins that conserve structural and active site features that can be explicitly associated with a conserved partial reaction or other chemical capability. Within a superfamily and constrained by these superfamily-common features, many divergent families may have evolved that exhibit different reaction and/or substrate specificities (9). (See the Prologue for some definitions of superfamilies, families, and related terms.)

These types of superfamilies provide a useful context for inference of functional properties of members of unknown function ("unknowns") because the constraints imposed by the structure-function paradigm unique to each superfamily restrict the search space for functional inference of their reaction and substrate specificities, simplifying their functional assignments. Because the number of sequences in each superfamily is still increasing rapidly, large amounts of new data are regularly available to inform these investigations. Moreover, sequence and structural similarities among all of the members of a superfamily can be associated with many types of functional information, allowing us to leverage what is known to guide inference of functional properties of unknowns that are similar. (See the minireview by Gerlt *et al.* (48) in this thematic series describing strategies for assigning functions in the enolase superfamily for an example.) Furthermore, as our coverage of genome space increases, new "outlier" functions in superfamilies can be identified from specialized environmental niches, extending our estimates of the natural boundaries of functional variation that a particular superfamily supports.

Below, we describe how the continuing increase in sequence and structural data can be used to understand better the evolution of new functions and to improve functional inference accessed using a relatively new application of network-based methods, protein similarity networks, an attractive approach for investigation of functional properties from the context of sequence and structural similarity. Results from such large-scale studies are reviewed here using examples from three different superfamilies of enzymes: the eukaryotic protein kinase

* This work was supported, in whole or in part, by National Institutes of Health Grants R01 GM60595 and U54 GM093342. This is the fifth article in the Thematic Minireview Series on Enzyme Evolution in the Post-genomic Era.

[1] To whom correspondence should be addressed. E-mail: babbitt@cgl.ucsf.edu.

(ePK)[2]-like superfamily, a large group of acid-sugar dehydratases from the enolase superfamily, and the glutathione transferase (GST) superfamily.

Emerging Roles for Large-scale Computational Analysis of Protein Superfamilies

As methods for managing and analyzing sequence and structural data have improved, computational studies can more effectively address broad issues in large-scale mapping of structure-function relationships and deduction of the patterns by which natural evolution has led to the divergence of many functions from an ancestral structural scaffold. For example, for protein kinases, one of the largest and most important enzyme superfamilies, the seminal Manning tree (10) provided a foundation for classification of human kinases and those from other eukaryotes. Likewise, a large-scale study of redox proteins generated a census of sequence, structural, and functional characteristics of the divergent superfamilies of the thioredoxin fold class that are represented in nature (11).

Large-scale analyses have the additional advantage of revealing patterns not easily observable when smaller data sets are examined. For example, comparison of sequence and structural features conserved in the active sites of the members of the large and functionally diverse enolase superfamily allowed the prediction of the specific partial reaction uniting the entire superfamily, the abstraction of an α-proton of a carboxylic acid, thereby restricting the functional prediction problem for the thousands of sequences now identified as superfamily members to consideration of only the overall reactions and substrates consistent with that paradigm (12). Using that structure-function mapping as a foundation, more detailed computational and experimental studies have identified differences among superfamily members that distinguish the reaction and substrate specificities of the >20 constituent families whose functions can now be assigned (see the minireview by Gerlt *et al.* (48) for a listing). Other notable studies linking structural and mechanistic features across large enzyme superfamilies include analyses of the amidohydrolase (13, 14), enoyl-CoA hydratase (15), nudix (16), haloalkanoic acid dehalogenase (17), and two dinucleotide-binding domain flavoprotein (18) superfamilies, to name a few.

As more powerful tools and computers have been created, the ease of mounting such studies has enabled new types of analyses that provide context for interpreting functional characteristics across homologous members of superfamilies. These include sophisticated algorithms for multiple alignment and phylogenetic inference, both of which have long been used to examine evolutionary relationships among groups of sequences. Especially relevant to this minireview, phylogenomic approaches, first described over a decade ago (19), combine phylogenetic reconstruction with functional assignment of unknowns based on their placement in the tree relative to knowns. Phylogenomic approaches have now been applied extensively to improve the accuracy of homology-based annotation and to distinguish divergent families within enzyme superfamilies (see Ref. 20 for an example). Additionally, searchable online databases such as BRENDA (21) provide access to a large store of enzyme function information, whereas others provide online curation and computational tools created to link enzyme sequence and structural information with functional characteristics and mechanistic properties (22–25).

Network-based Approaches for Large-scale Analysis of Protein Superfamilies

Although large-scale analyses indeed provide a "big picture" perspective that adds much to our understanding of genomic and chemical biology, the growing size of the data sets and their associated metadata continue to raise significant challenges for analysis and dissemination. Network-based analysis represents one approach used to capture biological context, with genetic or protein interaction networks using computational and/or experimental data being among the most common. Sequence and structure similarity networks have also been used for the analysis and visualization of structure-function relationships (26–28). This technique allows users to efficiently and quickly examine similarities of much larger sets of proteins than is generally possible using traditional methods such as phylogenetic trees and multiple alignments. For example, one such study mounted a comparison of over 145,000 sequences to create a map in which proteins are positioned according to sequence relationships and gene functions (29). The recent development of software platforms such as Cytoscape (30) facilitates the use of network methods and algorithms of several types, enabling access to these types of tools by non-experts.

Although they are not a substitute for phylogenetic inference, networks generated from even such simple metrics as all-by-all pairwise comparisons of a large number of divergent sequences have been shown to track well with known relationships and with the clustering provided by trees. Furthermore, they support facile mapping of many types of orthogonal data to proteins clustered by similarity (31). Types of information such as genome/operon context, interaction networks and pathways, and organism-specific information have been shown to enhance the accuracy of functional inference (see Refs. 32 and 33 for relevant reviews). In analogy to phylogenomics, functional information of many types can be associated with nodes (*e.g.* protein sequences or structures) in a similarity network to improve functional inference and insight. Because protein similarity networks can be quickly generated in interactive formats, users can easily explore these associations by coloring nodes with different combinations of sequence/structural properties and functional information.

Examples illustrating the application of large-scale analysis of structure-function relationships using protein similarity networks are described below. Interactive versions of these networks are available from the authors and can be viewed using the freely available Cytoscape software (30).

Tracking Growth of Structural Coverage: ePK-like Superfamily

The ePK-like superfamily is a large and diverse group of homologous enzymes that share a common protein kinase-like

[2] The abbreviations used are: ePK, eukaryotic protein kinase; GST, glutathione transferase; HMM, hidden Markov model; SFLD, Structure-Function Linkage Database; r.m.s.d., root mean square deviation.

fold (34) and conserved residues associated with ATP-dependent phosphorylation of proteins and small molecules. ePK-like enzymes mediate many important cellular processes, including signal transduction (10). They make up almost 2% of eukaryotic genes and, although present as a smaller percentage of bacterial genes, may be at least as important in bacterial cellular regulation as the structurally unrelated histidine kinases (35).

The size and diversity of the ePK-like superfamily make it hard to generate a global overview of their sequence and structural relationships. As a result, only a small number of groups have attempted the time-consuming task of generating large-scale classifications of the kinases. In one of these studies, Kannan *et al.* (35) used a library of hidden Markov models (HMMs) to identify >45,000 ePK-like sequences from the NCBI nonredundant database (36) and the Global Ocean Sampling data set (37) and to classify them into 20 families. Examination of this diverse sequence set allowed the identification of 10 residues conserved across most families. Six of these residues were known to be involved in ATP and substrate binding and catalysis, whereas the functional role of the remaining residues had not been established. This study also showed that all but one of these well conserved residues had been lost over the course of evolution in one or more families (in some cases, substituted with changes in other regions of the protein), illustrating the plasticity of the ePK-like fold. Although profile-profile alignments and alignments of conserved motifs could be used to group some families into related clusters, the size and diversity of the superfamily have continued to challenge the construction of a more detailed evolutionary history.

Scheeff and Bourne (38) were able to surmount the problem of low sequence identity across the superfamily by combining sequence and structural information into a single phylogenetic analysis. The results suggested that the tree constructed by this method had some advantages and was more reliable than trees produced using either sequence or structural data alone.

In addition to these types of global analyses, many thousands of detailed studies have been published describing properties of smaller groups and of individual enzymes. However, the sheer number of sequences and structures in this superfamily, coupled with the rate of growth of the sequence and structure databases, makes keeping an up-to-date record of kinase relationships increasingly difficult, even without the inclusion of linked functional information. (The Pfam (39) PKinase clan currently includes nearly 85,000 sequences.) Here, we illustrate the use of similarity networks to keep track of relationships between enzymes in large superfamilies. In this example, networks generated from pairwise structural comparisons provide a current update of the structural coverage of the superfamily.

Fig. 1 shows structure similarity networks for the ePK-like superfamily,[3] colored by Pfam classifications, with Fig. 1*A* indicating the differences in structural coverage in the years between when the study by Scheeff and Bourne (38) was published (October 2005) and May 2011, respectively. As is clear from these summaries, the structure space has filled out significantly over this 6-year span. Most strikingly, the fructosamine kinase family defined by Pfam, Fructosamin_kin (*red oval* in Fig. 1*A*, *lower panel*), was not represented at all in the network from 2005. Fig. 1*B* shows the same network as in Fig. 1*A* (*lower panel*), but thresholded at a higher stringency scoring cutoff (achieved by increasing the score threshold required for drawing edges between two nodes), enabling a more detailed view of the same structural relationships. Fig. 1*B* provides a different and somewhat more detailed view of the growth of structural coverage between these two time points. Although these networks use a set of structures that is larger and somewhat different from that used by Scheeff and Bourne, they track reasonably well with those trees (data not shown). Some exceptions include structures for which the position was labeled as uncertain in the Scheeff and Bourne tree. Alternative versions of these networks colored by the Manning classification (10), with the addition of the atypical kinase class used in Ref. 38, are provided in Fig. 2.

As shown in this example, similarity networks can be used effectively to update relationships among proteins in a superfamily as new structures become available, if, as for the ePK-like superfamily, its structural coverage is good. Sequence networks can also be used to summarize relationships among proteins on a large scale (11), as described below. Although the scale at which networks can easily query such data is still much larger than can generally be accommodated using multiple alignments and trees, the size of networks that can be viewed and manipulated by software such as Cytoscape is limited by the number of edges they contain. In practice, for a superfamily as large as the kinases, only a small proportion of the available sequences can be represented in a single network, typically requiring the use of representative sequences to cover the divergence space. Additionally, because of the diversity of many superfamilies, including the ePK-like superfamily, it is not possible to connect the whole set of sequences at statistically significant scores.

Prediction of New Carbon Sources in Human Gut Microbiome from Comparisons with Acid-sugar Dehydratases of Enolase Superfamily

Microbes residing in the gut have a significant influence on human health. In addition to aiding in energy harvest from food and synthesizing essential vitamins, changes in the gut microbial population are associated with medical conditions such as inflammatory bowel disease and obesity (3). Variations in microbiome populations have also been observed following treatment with antibiotics (40). Thus, much interest is now focused on determining the molecular functions and biological roles of the gut metaproteome both in healthy individuals and in those suffering from disease.

One of the most comprehensive studies on the human gut microbiome to date describes a set of 3.3 million microbial genes sequenced and assembled from fecal samples of 124 individuals (3). As expected, the census of protein functions initially identified in this metagenome includes proteins in many cen-

[3] For network analysis for the ePK-like superfamily, structures were chosen to include only one structure for each unique UniProt ID, with a preference for 1) structures solved October 2005 or previously and 2) wild-type, 3) ligand-bound, and 4) good resolution structures. Using the FAST algorithm (46), each structure in the set was used as a query against a database containing all structures in the set. Networks were created at various *N*-score cutoffs and visualized using Cytoscape.

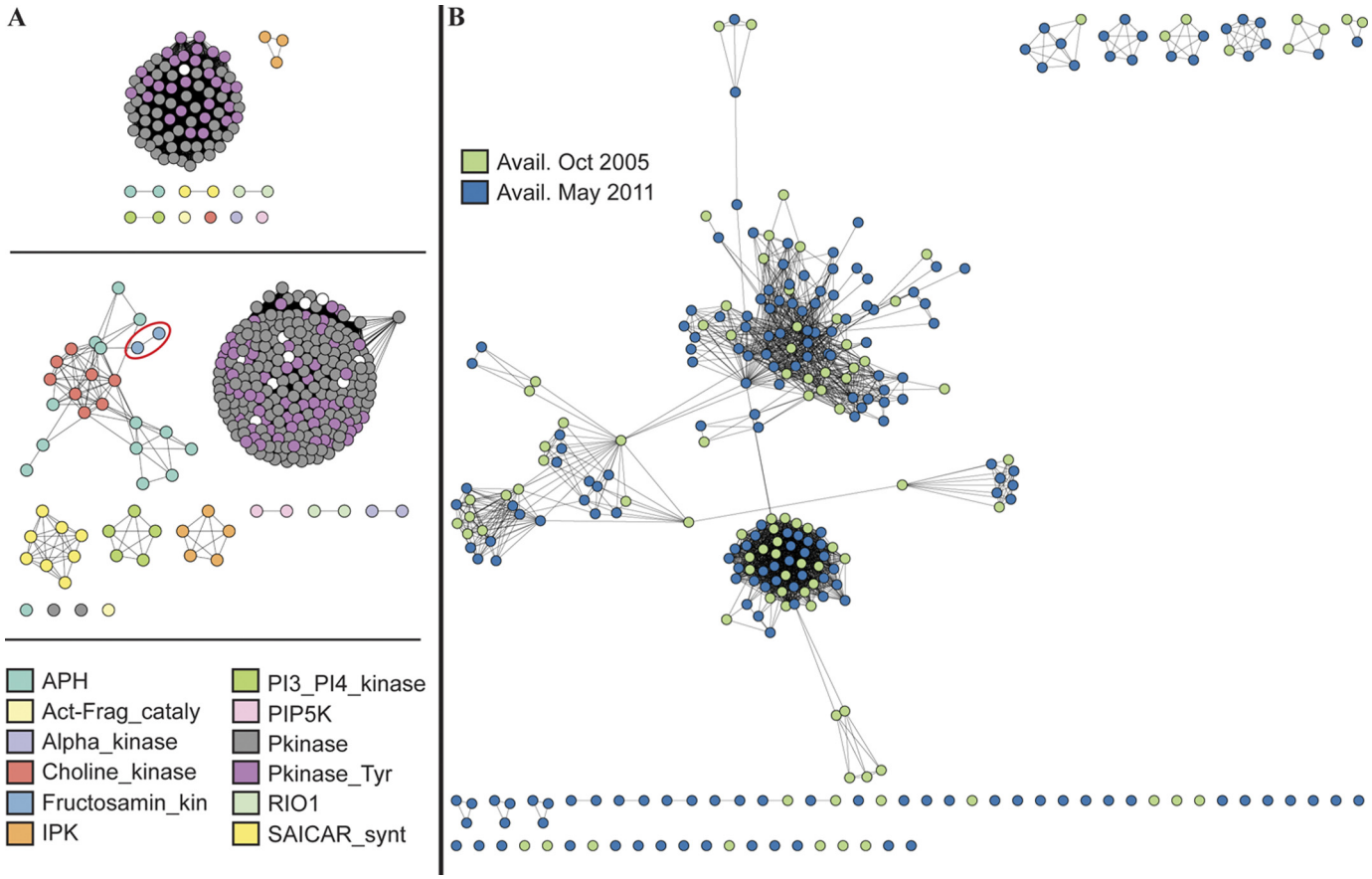

FIGURE 1. **Structure similarity networks of ePK-like superfamily generated from pairwise comparisons using FAST algorithm.** Each node represents a structure. Each edge represents a connection with a FAST *N*-score better than a given threshold. *A*, FAST *N*-score cutoff = 11, colored by Pfam family. *Upper panel*, structures available as of October 2005 (97 nodes). At this cutoff, the average root mean square deviation (r.m.s.d.) is ~2.81 Å with ~213 Cα atoms aligned. *Lower panel*, structures available as of May 2011 (295 nodes). At this cutoff, the average r.m.s.d. is ~2.98 Å with ~207 Cα atoms aligned. *B*, FAST *N*-score cutoff = 23. At this cutoff, the average r.m.s.d. is ~1.97 with ~247 Cα atoms aligned. Nodes colored *green* represent structures available in the Protein Data Bank as of October 2005; those colored *blue* represent structures added to the Protein Data Bank between October 2005 and May 2011 (total of 295 nodes). Nodes were arranged using the yFiles organic layout provided with Cytoscape version 2.7. Lengths of edges are not meaningful except that sequences in tightly clustered groups are relatively more similar to each other than sequences with few connections.

tral metabolic pathways such as those involved in carbon utilization pathways. We used the information available in the Structure-Function Linkage Database (SFLD)[4] (25) for a large set of acid-sugar dehydratases in the enolase superfamily to probe for additional and possibly unique carbon sources in the microbiome. This was accomplished by identifying putative acid-sugar dehydratases in the gut metagenome that differ from those that had been previously identified, whether of known or unknown specificity.

The substrate specificities of 10 acid-sugar dehydratases have now been biochemically established,[5] allowing functional assignment of specificity to ~40% of the ~2000 sequences currently represented in this subgroup of the superfamily in SFLD. Although the rest can be assigned with high confidence as likely acid-sugar dehydratases, their substrate specificities remain unknown. Using SFLD tools, protein sequences from the human gut microbiome predicted to be acid-sugar dehydratases were identified and clustered together with the knowns and unknowns of the subgroup already annotated in the database. The results are summarized in the network shown in Fig. 3*A*.[6]

This network is thresholded at a relatively permissive cutoff, where most families are found in one major cluster. Other reaction families that do not show similarities to any of the nodes in

[4] SFLD is a joint project of the Babbitt laboratory (supported by National Institutes of Health Grant GM60595 and National Science Foundation Grants DBI-0234768 and DBI-0640476) and the UCSF Resource for Biocomputing, Visualization, and Informatics (supported by National Institutes of Health Grant P41 RR001081). Additional support for the creation of networks available at SFLD is provided by the Enzyme Function Initiative (supported by National Institutes of Health Grant U54 GM093342).

[5] Of 10 acid-sugar dehydratase families of known reaction specificity in SFLD, only seven are colored in Fig. 3, as two others are not represented in this analysis. The mandelate racemase family, the namesake of the subgroup, is also colored. Although mandelate racemase is not an acid-sugar dehydratase, it is a member of this subgroup by sequence and structural similarity and is therefore included in Fig. 3.

[6] For network analysis for the gut metagenome, the sequence set consists of 1) the subgroup from SFLD containing acid-sugar dehydratases (named the mandelate racemase subgroup), filtered to 90% identity, aside from experimentally characterized members, all of which are present, and 2) all gut metagenome sequences that matched either this SFLD subgroup HMM or an SFLD family HMM from a family within the subgroup with an *e*-value cutoff of at least 1*e*−2 and that did not better match any other enolase superfamily SFLD HMMs. These sequences were filtered to 90% identity and to remove fragments under 150 amino acids. BLAST analysis (47) was performed using each sequence in the set as a query against a database containing all sequences in the set. Networks were created at two different *e*-value cutoffs and visualized as described in Footnote 3.

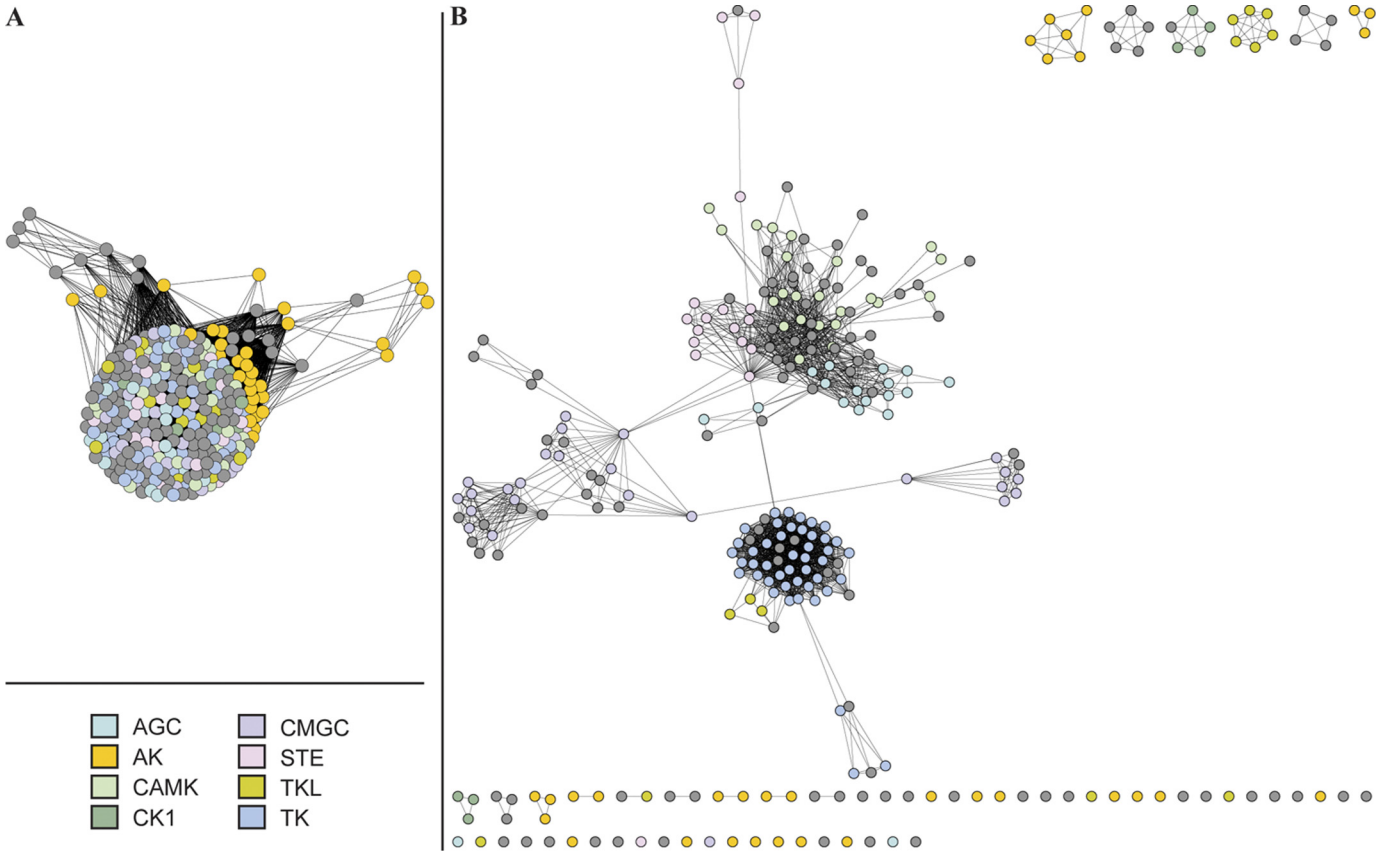

FIGURE 2. **Alternative view of structure similarity networks of 86 representative structures in ePK-like superfamily (generated as described for Fig. 1).** Nodes are colored according to their Manning/Bourne group classification. *Dark gray nodes* represent structures that were not classified. *A*, FAST *N*-score cutoff = 4. *B*, FAST *N*-score cutoff = 23.

this large cluster at a threshold better than the cutoff form smaller clusters arranged randomly at the bottom of Fig. 3*A*. Simple examination reveals a few emerging clusters in the main cluster and also in the separated clusters (*e.g.* the *circled* group in Fig. 3*A*) that are populated primarily or exclusively by gut metagenomic sequences. Because these sequences are somewhat distant from those with characterized functions (designated by different colors), they may indeed represent unique acid-sugar dehydratases and, hence, new carbon sources not previously associated with the superfamily.

A more detailed examination of this hypothesis can be obtained by visualization of the network at the more stringent *e*-value cutoff, shown in Fig. 3*B*. In this view, most of the characterized families within the subgroup have separated into individual clusters, suggesting that this threshold cutoff may be useful for hypothesizing the boundaries of at least some of the functionally distinct families within it. From this view, we can predict the specificity of some of the metagenomic sequences that cluster closely with known families, *e.g.* fuconate and galactonate dehydratases. The perspective provided in Fig. 3*B* also lends support to the hypothesis that the separated clusters populated only by gut metagenomic sequences and other uncharacterized sequences from the GenBank™ Data Bank may indeed represent new carbon sources not previously identified as members of the enolase superfamily. Finally, the addition of these metagenomic sequences to the networks helps to fill out the sequence space representing the acid-sugar dehydratases and illustrates more fully the breadth of their natural diversity. It is also interesting that some clusters containing members of characterized families in Fig. 3*B* have no representatives from the gut microbiome, suggesting that these functions may not be represented in the microorganisms that live in the gut (or those functions are supplied by enzymes from a different evolutionary background).

What We Do Not Know About Cytosolic GST Superfamily

GSTs constitute a large class of enzymes that play important biological roles in cell signaling and metabolism of endogenous compounds, drugs, and other xenobiotics. They are ubiquitous in nature (except for archaea) and may represent as much as 0.01% of the enzyme universe.[7] Based on sequence similarities, GSTs have historically been organized into major classes using the names of Greek letters (*e.g.* Alpha, Pi, Omega, Theta, etc.) (41). Within each major class, subclasses designate functional and other properties. Although a number of GSTs have been experimentally characterized in terms of their general substrate profiles, the physiological substrates and reaction specificities of only a small minority are known. Still, because of their importance to human biology and health, GSTs are among the best studied of enzyme superfamilies, with thousands of publications detailing their biological roles and structural and functional properties.

[7] H. J. Atkinson and P. C. Babbitt, unpublished data.

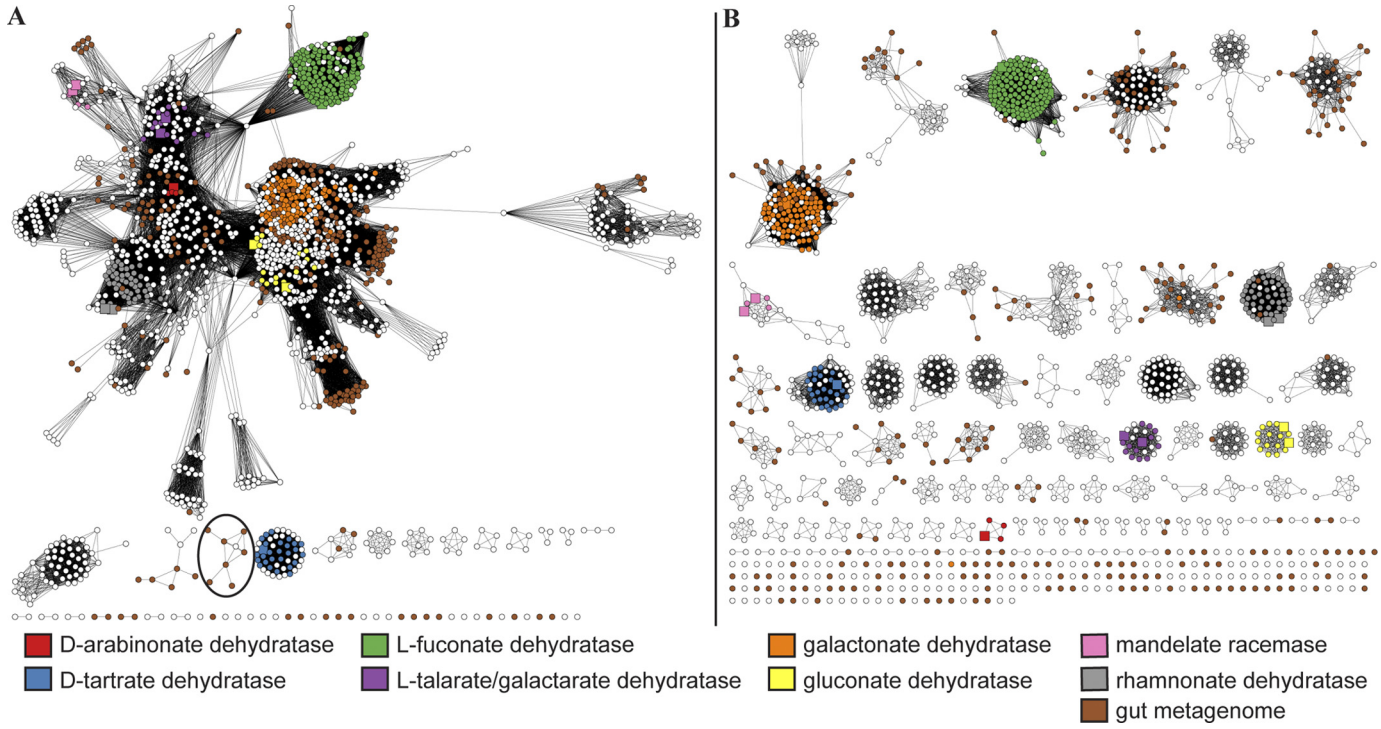

FIGURE 3. **Sequence similarity networks of acid-sugar dehydratases known or predicted to belong to enolase superfamily and human gut microbiome.** Networks were generated from all-by-all BLAST comparisons of 1578 sequences representing sequences of eight known acid-sugar dehydratase families and the mandelate racemase family from the mandelate racemase subgroup (see Footnote 5) as defined by SFLD and a filtered set of gut metagenome sequences that showed significant similarity to the members of the subgroup. Each of the 1578 nodes represents a sequence. *Larger square nodes* represent those that have been experimentally characterized, so their reaction and substrate specificities are known. *Brown nodes* represent sequences from the human gut metagenome, and *white nodes* represent SFLD sequences in the subgroup for which the reaction and substrate specificities have not been predicted. The remainder (*small nodes*) represent sequences for which specificity can be predicted at high confidence, colored by their SFLD family names (see Footnote 4). Nodes were arranged using the yFiles organic layout provided with Cytoscape version 2.7. *A*, each edge in the network represents a BLAST connection with an *e*-value of 1*e*−44 or better. At this cutoff, sequences have a median percent identity and alignment length of ~32% and 369, respectively. *B*, each edge in the network represents a BLAST connection with an *e*-value of 1*e*−84 or better. At this cutoff, sequences have a median percent identity and alignment length of ~44% and 384, respectively. Lengths of edges are not meaningful except that sequences in tightly clustered groups are relatively more similar to each other than sequences with few connections.

Only a few studies have focused on the GST superfamily on a large scale, however (11, 42, 43). The sequence similarity network[8] shown in Fig. 4 provides an overview of the cytosolic GST superfamily from one of these (42). It compares 622 GSTs representing >6000 sequences and shows that they can be divided into two major groups distinguished by sequence and structural similarity (and also by variations in their active site features). The majority of the enzymes in the smaller of the two groups shown in Fig. 4 (*Group 1*) are from eukaryotic organisms, whereas those from the larger group (*Group 2*) are more mixed, but with the largest number coming from bacteria.

The summary of sequence relationships and structural coverage provided in Fig. 4 is the first time that similarity relationships across the entire GST superfamily were captured in a single view. This map shows both the sequences that could be classified as members of one of the major classes (*colored nodes*) as well as those that had not even been assigned to one of these general classes (*light* and *dark gray nodes*) and had thus far only been identified as belonging to the cytosolic GST superfamily. Remarkably, despite decades of study, these results reveal that the huge majority of GSTs have never been functionally characterized at any level. Furthermore, the representation of the colored nodes in the overall topology suggests that many additional classes likely remain to be defined. The view provided in Fig. 4 thus lays a foundation for choosing new sequences for which functional and structural characterization may be especially valuable for prediction of new functional classes. Many additional GST sequences have recently been identified,[9] so the proportion of GSTs for which no functional information is available continues to increase dramatically.

Challenges for Computational Prediction of Functional Properties

The examples provided in this minireview suggest the value of large-scale analyses such as similarity networks for summarizing sequence and structural relationships in large superfamilies and for developing hypotheses about how structure- or sequence-based clustering tracks with functional boundaries. However, like any other method, similarity networks also have some significant limitations, a few of which have been addressed above and others elsewhere (31). Although it is only by experimental investigation that the *in vitro* and *in vivo* functions of unknowns can ultimately be validated, the continual

[8] For network analysis for the GST superfamily, the sequence set was generated, and networks were calculated and visualized as described previously (42).

[9] P. C. Babbitt and D. Stryke, unpublished data.

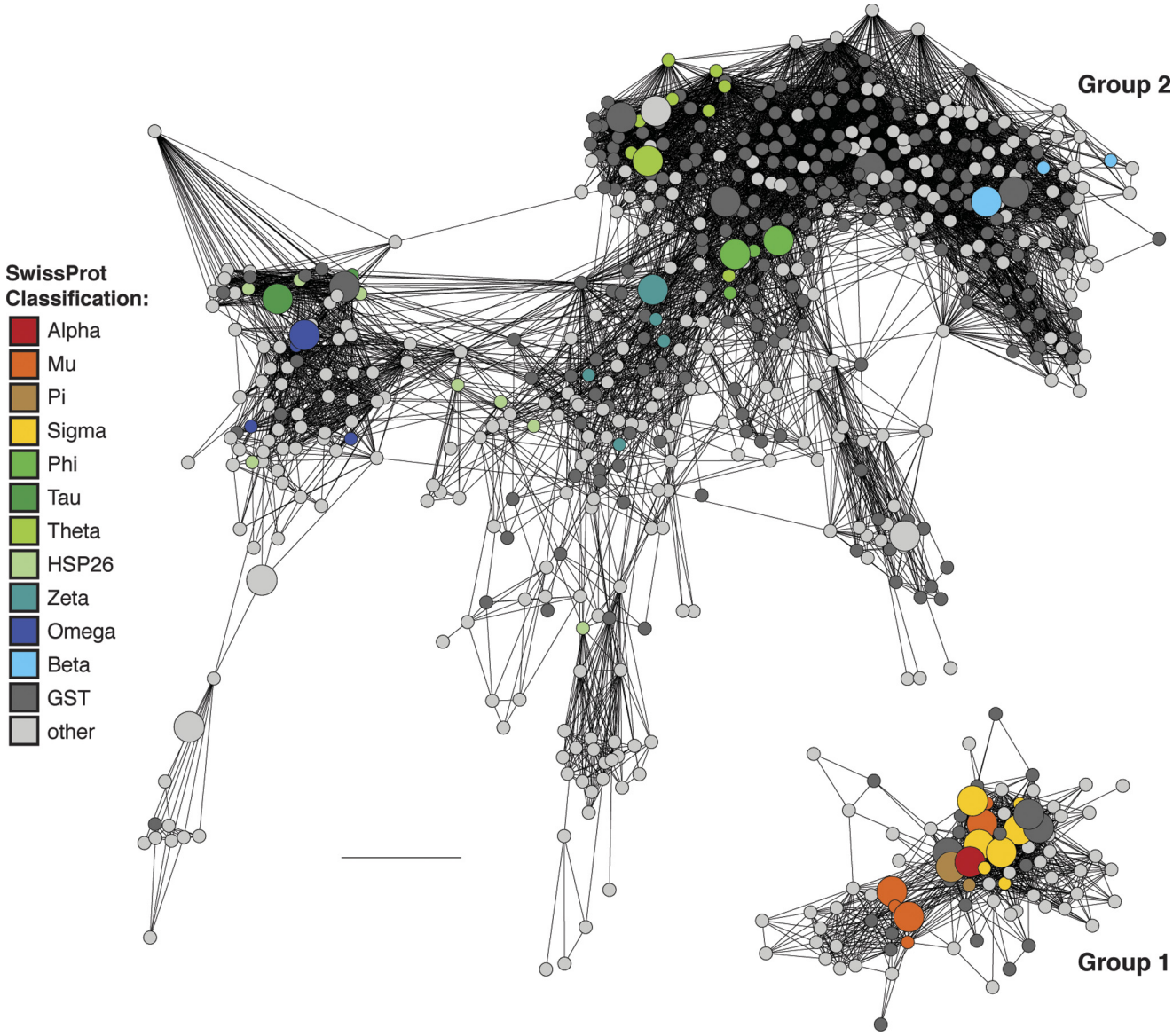

FIGURE 4. **Sequence similarity network of cytosolic GSTs.** Similarity is defined by pairwise BLAST alignments better than an *e*-value cutoff of 1*e*−12. 622 representative sequences that are a maximum of 40% identical and that span the diversity of >6000 GSTs are shown. Nodes are colored by classification of the sequence in the Swiss-Prot Database (part of the UniProt Database), if available. The *40 large nodes* designate sequences with structures. At this cutoff, edges at this threshold represent alignments with a median 27% identity over 200 residues. This network and legend are adapted from Ref. 42 with permission.

growth of sequence data makes it increasingly difficult for either focused or high-throughput experimental studies to keep up. Even a reasonable fallback position requires the development of new strategies for identifying the few experiments that could be most useful for validation of large-scale computational predictions. As illustrated here and elsewhere (44, 45), protein similarity networks represent one way to generate the context needed for choosing those experiments and interpreting the results.

REFERENCES

1. UniProt Consortium (2011) *Nucleic Acids Res.* **39,** D214–D219
2. Dutta, S., Burkhardt, K., Young, J., Swaminathan, G. J., Matsuura, T., Henrick, K., Nakamura, H., and Berman, H. M. (2009) *Mol. Biotechnol* **42,** 1–13
3. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J. M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. (2010) *Nature* **464,** 59–65
4. Roberts, R. J., Chang, Y. C., Hu, Z., Rachlin, J. N., Anton, B. P., Pokrzywa, R. M., Choi, H. P., Faller, L. L., Guleria, J., Housman, G., Klitgord, N., Mazumdar, V., McGettrick, M. G., Osmani, L., Swaminathan, R., Tao, K. R., Letovsky, S., Vitkup, D., Segrè, D., Salzberg, S. L., Delisi, C., Steffen, M., and Kasif, S. (2011) *Nucleic Acids Res.* **39,** D11–D14
5. Bateman, A. (2010) *Bioinformatics* **26,** 991
6. Raes, J., Harrington, E. D., Singh, A. H., and Bork, P. (2007) *Curr. Opin. Struct. Biol.* **17,** 362–369
7. Hsiao, T. L., Revelles, O., Chen, L., Sauer, U., and Vitkup, D. (2010) *Nat. Chem. Biol.* **6,** 34–40
8. Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) *PLoS Comput. Biol.* **5,** e1000605
9. Gerlt, J. A., and Babbitt, P. C. (2001) *Annu. Rev. Biochem.* **70,** 209–246

10. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) *Science* **298,** 1912–1934
11. Atkinson, H. J., and Babbitt, P. C. (2009) *PLoS Comput. Biol.* **5,** e1000541
12. Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., and Gerlt, J. A. (1996) *Biochemistry* **35,** 16489–16501
13. Holm, L., and Sander, C. (1997) *Proteins Struct. Funct. Genet.* **28,** 72–82
14. Seibert, C. M., and Raushel, F. M. (2005) *Biochemistry* **44,** 6383–6391
15. Holden, H. M., Benning, M. M., Haller, T., and Gerlt, J. A. (2001) *Acc. Chem. Res.* **34,** 145–157
16. Mildvan, A. S., Xia, Z., Azurmendi, H. F., Saraswat, V., Legler, P. M., Massiah, M. A., Gabelli, S. B., Bianchet, M. A., Kang, L. W., and Amzel, L. M. (2005) *Arch. Biochem. Biophys.* **433,** 129–143
17. Burroughs, A. M., Allen, K. N., Dunaway-Mariano, D., and Aravind, L. (2006) *J. Mol. Biol.* **361,** 1003–1034
18. Ojha, S., Meng, E. C., and Babbitt, P. C. (2007) *PLoS Comput. Biol.* **3,** e121
19. Eisen, J. A. (1998) *Genome Res.* **8,** 163–167
20. Brown, D. P., Krishnamurthy, N., and Sjölander, K. (2007) *PLoS Comput. Biol.* **3,** e160
21. Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., and Schomburg, D. (2011) *Nucleic Acids Res.* **39,** D670–D676
22. Gariev, I. A., and Varfolomeev, S. D. (2006) *Bioinformatics* **22,** 2574–2576
23. Holliday, G. L., Almonacid, D. E., Bartlett, G. J., O'Boyle, N. M., Torrance, J. W., Murray-Rust, P., Mitchell, J. B., and Thornton, J. M. (2007) *Nucleic Acids Res.* **35,** D515–D520
24. Nagano, N. (2005) *Nucleic Acids Res.* **33,** D407–D412
25. Pegg, S. C., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., Chang, P. J., Huang, C. C., Ferrin, T. E., and Babbitt, P. C. (2006) *Biochemistry* **45,** 2545–2555
26. Enright, A. J., and Ouzounis, C. A. (2001) *Bioinformatics* **17,** 853–854
27. Frickey, T., and Lupas, A. (2004) *Bioinformatics* **20,** 3702–3704
28. Huttenhower, C., Mehmood, S. O., and Troyanskaya, O. G. (2009) *BMC Bioinformatics* **10,** 417
29. Adai, A. T., Date, S. V., Wieland, S., and Marcotte, E. M. (2004) *J. Mol. Biol.* **340,** 179–190
30. Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P. L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., and Bader, G. D. (2007) *Nat. Protoc.* **2,** 2366–2382
31. Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) *PLoS ONE* **4,** e4345
32. Frishman, D. (2007) *Chem. Rev.* **107,** 3448–3466
33. Rentzsch, R., and Orengo, C. A. (2009) *Trends Biotechnol.* **27,** 210–219
34. Taylor, S. S., and Radzio-Andzelm, E. (1994) *Structure* **2,** 345–355
35. Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C., and Manning, G. (2007) *PLoS Biol.* **5,** e17
36. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011) *Nucleic Acids Res.* **39,** D32–D37
37. Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S., Li, H., Mashiyama, S. T., Joachimiak, M. P., van Belle, C., Chandonia, J. M., Soergel, D. A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B. J., Bafna, V., Friedman, R., Brenner, S. E., Godzik, A., Eisenberg, D., Dixon, J. E., Taylor, S. S., Strausberg, R. L., Frazier, M., and Venter, J. C. (2007) *PLoS Biol.* **5,** e16
38. Scheeff, E. D., and Bourne, P. E. (2005) *PLoS Comput. Biol.* **1,** e49
39. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010) *Nucleic Acids Res.* **38,** D211–D222
40. Dethlefsen, L., and Relman, D. A. (2011) *Proc. Natl. Acad. Sci. U.S.A.* **108,** 4554–4561
41. Mannervik, B., Board, P. G., Hayes, J. D., Listowsky, I., and Pearson, W. R. (2005) *Methods Enzymol.* **401,** 1–8
42. Atkinson, H. J., and Babbitt, P. C. (2009) *Biochemistry* **48,** 11108–11116
43. Pearson, W. R. (2005) *Methods Enzymol.* **401,** 186–204
44. Hicks, M. A., Barber, A. E. I., Giddings, L. A., Caldwell, J., O'Connor, S. E., and Babbitt, P. C. (2011) *Proteins Struct. Funct. Genet.* **79,** 3082–3098
45. Pieper, U., Chiang, R., Seffernick, J. J., Brown, S. D., Glasner, M. E., Kelly, L., Eswar, N., Sauder, J. M., Bonanno, J. B., Swaminathan, S., Burley, S. K., Zheng, X., Chance, M. R., Almo, S. C., Gerlt, J. A., Raushel, F. M., Jacobson, M. P., Babbitt, P. C., and Sali, A. (2009) *J. Struct. Funct. Genomics* **10,** 107–125
46. Zhu, J., and Weng, Z. (2005) *Proteins* **58,** 618–627
47. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402
48. Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., and Almo, S. C. (2012) *J Biol. Chem.* **287,** 29–34